Toward machines that behave ethically better than humans do

Pontier, M.A.; Hoorn, J.F.

published in
Proceedings of of the 34th International Annual Conference of the Cognitive Science Society
2012

document version
Peer reviewed version

Toward machines that behave ethically better than humans do

Pontier, M.A., Hoorn, J.F.

In: N. Miyake, B. Peebles, R.P. Cooper (Ed.), Proceedings of of the 34th International Annual Conference of the Cognitive Science Society (pp. 2198-2203). Austin, TX: Cognitive Science Society

Toward machines that behave ethically better than humans do

Matthijs A. Pontier (m.a.pontier@vu.nl)
Johan F. Hoorn (j.f.hoorn@vu.nl)
VU University Amsterdam, Center for Advanced Media Research Amsterdam (CAMeRA),
De Boelelaan 1081, 1081HV Amsterdam, The Netherlands

Abstract

With the increasing dependence on autonomous operating agents and robots the need for ethical machine behavior rises. This paper presents a moral reasoner that combines connectionism, utilitarianism and ethical theory about moral duties. The outcomes of the moral reasoning match those of expert ethicists in the health domain. This may be useful in many applications, especially where machines interact with humans in a medical context. Additionally, when connected to a cognitive model of emotional intelligence and affective decision making, it can be explored how moral decision making impacts affective behavior.

Keywords: Cognitive modeling, Machine ethics, Medical ethics

Introduction

In view of increasing intelligence and decreasing costs of artificial agents and robots, organizations increasingly use such systems for more complex tasks. With this development, we increasingly rely on the intelligence of agent systems. Because of market pressures to perform faster, better, cheaper and more reliably, this reliance on machine intelligence will continue to increase (Anderson, Anderson & Armen, 2005).

As the intelligence of machines increases, the amount of human supervision decreases and machines increasingly operate autonomously. These developments request that we should be able to rely on a certain level of ethical behavior from machines. As Rosalind Picard (1997) nicely puts it: ‘‘the greater the freedom of a machine, the more it will need moral standards’’. Especially when machines interact with humans, which they increasingly do, we need to ensure that these machines do not harm us or threaten our autonomy. This need for ethical machine behavior has given rise to a field that is variously known as Machine Morality, Machine Ethics, or Friendly AI (Wallach, Franklin & Allen, 2010).

There are many domains where machines could play a significant role in improving our quality of life as long as ethical concerns about their behaviors can be overcome (Anderson & Anderson, 2008). This may seem difficult, and incorporating ethical behavior into machines is indeed far from trivial. Moral decision making is arguably even one of the most challenging tasks for computational approaches to higher-order cognition (Wallach, Franklin & Allen, 2010).

Moreover, with the increasing complexity of autonomous agents and robots, it becomes harder to predict their behavior, and to conduct it along ethical guidelines. Some may argue that this is a good reason not to let machines be responsible for making ethical decisions. However, the behavior of machines is still far easier to predict than the behavior of humans. Moreover, human behavior is typically far from being morally ideal (Allen, Varner & Zinser, 2000). One of the reasons for this is that humans are not very good at making impartial decisions. We can expect machines to outperform us in this capability (Anderson & Anderson, 2010). Looking at it from this side, it seems that machines capable of sufficient moral reasoning would even behave ethically better than most human beings would. Perhaps interacting with ethical robots may someday even inspire us to behave ethically better ourselves.

There have been various approaches in giving machines moral standards, using various methods. One of them, called casuistry, looks at previous cases in which there is agreement about the correct response. Using the similarities with these previous cases and the correct responses to them, the machine attempts to determine the correct response to a new ethical dilemma.

Rzepka and Araki (2005) demonstrate an approach, in which their system learns to make ethical decisions based on web-based knowledge, to be ‘independent from the programmer’. They argue it may be safer to imitate millions of people, instead of a few ethicists and programmers. This seems useful for imitating human ethical behavior, but it does not seem plausible that machines using this method will be able to behave ethically better than humans. After all, the system bases its decision on the average behavior of humans in general, misbehavior included.

Guarini (2006) offers another approach that could be classified as casuistry. The presented system learns from training examples of ethical dilemmas with a known correct response using a neural network. After the learning process, it is capable of providing plausible responses to new ethical dilemmas. However, reclassification of cases remains problematic in his approach due to a lack of reflection and explicit representation. Therefore, Guarini concludes that casuistry alone is not sufficient.

Anderson and Anderson (2007) agree to this conclusion, and address the need for top-down processes. The two most dominant top-down mechanisms are (1) utilitarianism and (2) ethics about duties. Utilitarians claim that ultimately morality is about maximizing the total amount of ‘utility’ (a measure of happiness or well being) in the world. The competing ‘big picture’ view of moral principles is that ethics is about duties and, on the flip side of duties, the rights of individuals (Wallach, Allen & Smit, 2008).

The two competitors described above may not differ as much as it seems. Ethics about duties can be seen as a useful model to maximize the total amount of utility. Thinking about maximizing the total amount of utility in a too direct manner may lead to a sub-optimal amount of utility. For example, in the case of the decision to kill one person to save five, killing the one person seems to maximize the total amount of utility. After all, compared to the decision of inaction, it leads to a situation with four more survivors (Anderson, Anderson & Armen, 2006). However, for humans it may be impossible to favor the decision of killing a person in this case over the decision of inaction, without

also making it more acceptable in other cases to kill human beings. Therefore, not having the intuition that it is wrong to kill one person to save more people would probably lead to a smaller total amount of utility in the world.

Anderson, Anderson and Armen (2006) use Ross's prima facie duties (Ross, 1930). Here, prima facie means a moral duty may be overruled by a more pressing one. They argue that the ideal ethical theory incorporates multiple prima facie duties with some sort of a decision procedure to determine the ethically correct action in cases where the duties give conflicting advice. Their system learns rules from examples using a machine learning technique. After learning, the system can produce correct responses to unlearned cases.

However, according to Wallach, Franklin and Allen (2010), the model of Anderson, Anderson and Armen (2006) is rudimentary and cannot accommodate the complexity of human decision making. In their work, Wallach et al. make a distinction between top-down and bottom-up moral-decision faculties and present an approach that combines both directions. They argue that the capacity for moral judgment in humans is a hybrid of both bottom-up mechanisms shaped by evolution and learning, and top-down mechanisms capable of theory-driven reasoning. Morally intelligent robots will eventually need a similar fusion, which maintains the dynamic and flexible morality of bottom-up systems, which accommodate diverse inputs, while subjecting the evaluation of choices and actions to top-down principles that represent ideals we strive to meet. Wallach, Franklin & Allen (2010) explore the possibility to implement moral reasoning in LIDA, a model of human cognition. This system combines a bottom-up collection of sensory data, such as in the neural network approach of Guarini (2006), with top-down processes for making sense of its current situation, to predict the results of actions. However, the proposed model is not fully implemented yet.

The current paper can be seen as a first attempt in combining a bottom-up and top-down approach. It combines a bottom-up structure with top-down knowledge in the form of moral duties. It balances between these duties and computes a level of morality, which could be seen as an estimation of the influence on the total amount of utility in the world.

Wallach, Franklin and Allen (2010) argue that even agents who adhere to a deontological ethic or are utilitarians may require emotional intelligence as well as other ''supra-rational'' faculties, such as a sense of self and a theory of mind (ToM). Moreover, according to Tronto (1993), care is only thought of as good care when it is personalized. Therefore, we represented the system in such a way that it is easy to connect to Silicon Coppélia (Hoorn, Pontier and Siddiqui, 2011), a cognitive model of emotional intelligence and affective decision making. Silicon Coppélia contains a feedback loop, by which it can learn about the preferences of an individual patient, and personalize its behavior. Silicon Coppélia estimates an Expected Satisfaction of possible actions, based on bottom-up data combined with top-down knowledge. This compares to the predicted results of actions in Wallach, Franklin and Allen (2010).

For simulation purposes, we focus on biomedical ethics, because in this domain relatively much consensus exists about ethically correct behavior. There is an ethically defensible goal (health), whereas in other areas (such as business and law) the goal may not be ethically defensible (money, helping a 'bad guy') (Anderson & Anderson, 2007). Moreover, due to a foreseen lack of resources and healthcare personnel to provide a high standard of care in the near future (WHO, 2010), robots are increasingly being used in healthcare.

Healthcare is a valid case where robots genuinely contribute to treatment. For example, previous research showed that animal-shaped robots can be useful as a tool for occupational therapy. Robins et al. (2005) used mobile robots to treat autistic children. Further, Wada and Shibata (2007) developed Paro, a robot shaped like a baby-seal that interacts with users to encourage positive mental effects. Interaction with Paro has been shown to improve users' moods, making them more active and communicative with each other and caregivers. Research groups have used Paro for therapy at eldercare facilities and with those having Alzheimer's disease (Kidd, Taggart & Turkle, 2006; Marti et al., 2006). Banks, Willoughby and Banks (2008) showed that animal-assisted therapy with an AIBO dog helped just as good for reducing loneliness as therapy with a living dog.

By providing assistance during care tasks, or fulfilling them, robots can relieve time for the many duties of care workers. However, care robots require rigorous ethical reflection to ensure that their design and introduction do not impede the promotion of values and the dignity of patients at such a vulnerable and sensitive time in their lives (Van Wynsberghe, 2012)

According to Gillon (1994), beneficence, non-maleficence, autonomy and justice are the four basic prima facie moral commitments. Here, confidentiality and truthfulness can be seen as a part of autonomy. Because we aim to match the expert data given from Buchanan and Brock (1989), who focus on dilemmas between autonomy, beneficence and non-maleficence, we focus on these three moral duties in the remainder of this paper.

The moral reasoner and its relation to Silicon Coppélia

Silicon Coppélia (Hoorn et al., 2011) is a model of emotional intelligence and affective decision making. In this model, the agent perceives the user on several dimensions, which leads to feelings of involvement and distance. These feelings represent the affective component in the decision making process. The rational component consists of the expected utility of an action for the agent itself (i.e., the belief that an action leads to achieving desired goals).

The system contains a library of goals and each agent has a level of ambition for each goal. There are desired and undesired goals, all with several levels of importance. The levels of ambition the agent attaches to the goals are represented by a real value between [-1, 1], where a negative value means that the goal is undesired and a positive value means that the goal is desired. A higher value means that the goal is more important to the agent.

The system contains a library of actions from which the agents can perform. The agent has beliefs about actions inhibiting or facilitating goals, represented by a real value between [-1, 1], -1 being full inhibition, 1 being full facilitation.

The expected utilities of possible actions are calculated by looking at the goal-states it influences. If an action or a feature is believed to facilitate a desired goal or inhibits an undesired goal, this will increase its expected utility and vice versa. The following formula is used to calculate the expected utility for the agent itself.

ExpectedUtility(Action, Goal) =
Belief(facilitates(Action, Goal)) * Ambition(Goal)

Given the level of ambition for a goal and the believed facilitation of that goal by an action, the agent calculates the expected utility for itself of performing that action regarding that goal by multiplying the believed facilitation of the goal with the level of ambition for the goal.

In the current moral reasoner, the agent tries to maximize the total amount of utility for everyone. In complex situations, it would take too much computational load to calculate all possible consequences of an action for everyone, and extract this into a single value of 'morality' of the action. Therefore, the agent tries to estimate the morality of actions by following three moral duties. These three duties consist of seeking to attain three moral values: (1) Autonomy, (2) Non-Maleficence and (3) Beneficence. In the moral reasoner, the three duties are seen as 'moral goals' to satisfy everyone's needs as much as possible. This corresponds with Super's conceptualization of the relationship between needs and values: "values are objectives that one seeks to attain to satisfy a need" (Super, 1973). The moral reasoner aims to pick actions that serve these moral goals best.

What priorities should be given to these three moral goals? According to Anderson and Anderson (2008), the following consensus exists in medical ethics. A healthcare worker should challenge a patient's decision only if the patient is not capable of fully autonomous decision making (e.g., the patient has irrational fears about an operation) and there is either a violation of the duty of non-maleficence (e.g., the patient is hurt) or a *severe* violation of the duty of beneficence (e.g., the patient rejects an operation that will strongly improve his or her quality of life). In other words, Autonomy is the most important duty. Only when a patient is not fully autonomous, the other moral goals come into play. Further, Non-maleficence is a more important duty than Beneficence, because only a severe violation of Beneficence requires challenging a patient's decision, while *any* violation of Non-maleficence does. Therefore, the ambition level for the moral goal 'Autonomy' was set to the highest value and 'Non-maleficence', which was set to a higher value than the ambition level for 'Beneficence'. The ambition levels that were given to the moral goals in the moral reasoner can be found in Table 1.

The agent calculates estimated level of Morality of an action by taking the sum of the ambition levels of the three moral goals multiplied with the beliefs that the particular actions facilitate the corresponding moral goals. When

Table 1: Ambition levels for moral goals

Moral Goal	Ambition level
Non-Maleficence	0.74
Beneficence	0.52
Autonomy	1

moral goals are believed to be better facilitated by a moral action, the estimated level of Morality will be higher. . The following formula is used to calculate the estimated Morality of an action:

Morality(Action) =
Σ_{Goal}(Belief(facilitates(Action, Goal)) * Ambition(Goal))

Note that this is similar to calculating the Expected Utility in Silicon Coppélia. To ensure that the decision of a fully autonomous patient is never questioned, we added the following rule to the moral reasoner:

IF belief(facilitates(Action, autonomy) = max_value
THEN Moralilty(Action) = Morilaity(Action) + 2

As can be seen Figure 1, this can be represented as a simple neural network, where moral goals are associated with the possible actions via the belief strengths that these actions facilitate the three moral goals. A decision function F adds the rule and picks the action with the highest activation as output.

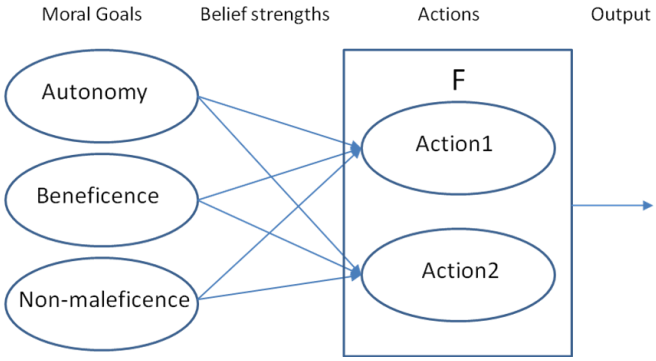

Figure 1: Moral reasoner shown in graphical format

Simulation Results

To see whether the moral reasoner could simulate the moral decision making of experts in medical ethics, the analysis of ethical dilemmas by expert ethicists was taken from Buchanan and Brock (1989). The following simulation experiments examine whether the moral reasoner reaches the same conclusions as these expert ethicists.

Experiment 1

In the simulated situation, the patient refuses to take an antibiotic that is almost certain to cure an infection that would otherwise likely lead to his death. The decision is the result of an irrational fear the patient has of taking medications. (For instance, perhaps a relative happened to die shortly after taking medication and this patient now believes that taking any medication will lead to death.)

According to Buchanan and Brock (1989), the correct answer is that the health care worker should try again to

change the patient's mind because if she accepts his decision as final, the harm done to the patient is likely to be severe (his death) and his decision can be considered as being less than fully autonomous.

As can be seen in Table 2, the moral reasoner also classifies the action 'Try again' as having a higher level of morality than accepting the decision of the patient. In this and the following tables, the fields under the three moral goals represent the believed facilitation of the corresponding moral goal by an action, as taken from Buchanan and Brock (1989). 'Non-Malef' stands for Non-maleficence, and 'Benef' stands for Beneficence.

Table 2: Simulation results of Experiment 1.

	Autonomy	**Non-Malef**	**Benef**	**Morality**
Try Again	**-0.5**	**1**	**1**	**0,76**
Accept	**0.5**	**-1**	**-1**	**-0,8**

Experiment 2

Once again, the patient refuses to take an antibiotic that is almost certain to cure an infection that would otherwise likely lead to his death, but this time the decision is made on the grounds of long-standing religious beliefs that do not allow him to take medications.

The correct answer in this case, state Buchanan and Brock (1989), is that the health care worker should accept the patient's decision as final because, although the harm that will likely result is severe (his death), his decision can be seen as being fully autonomous. The health care worker must respect a fully autonomous decision made by a competent adult patient, even if she disagrees with it, since the decision concerns his body and a patient has the right to decide what shall be done to his or her body.

As can be seen in Table 3, the moral reasoner comes to the correct conclusion. Here, the rule to ensure the decision of a fully autonomous patient is never questioned made a difference. If the rule would not have existed, the morality of 'Accept' would have been -0.3, and the moral reasoner would have concluded that it was more moral to try again.

Table 3: Simulation results of Experiment 2.

	Autonomy	**Non-Malef**	**Benef**	**Morality**
Try Again	**-0.5**	**1**	**1**	**0,76**
Accept	**1**	**-1**	**-1**	**1,7**

Experiment 3

Table 4: Simulation results of Experiment 3.

	Autonomy	**Non-Malef**	**Benef**	**Morality**
Try Again	**-0.5**	**0.5**	**0.5**	**0,13**
Accept	**1**	**-0.5**	**-0.5**	**2,37**

The patient refuses to take an antibiotic that is likely to prevent complications from his illness, complications that are not likely to be severe, because of long-standing religious beliefs that do not allow him to take medications.

The correct answer is that the health care worker should accept his decision, since once again the decision appears to be fully autonomous and there is even less possible harm at stake than in Experiment 2. The moral reasoner comes to the correct conclusion and estimates the Morality of 'Accept' higher than 'Try Again', as can be seen in Table 4

Experiment 4

A patient will not consider taking medication that could only help to alleviate some symptoms of a virus that must run its course. He refuses the medication because he has heard untrue rumors that the medication is unsafe.

Even though the decision is less than fully autonomous, because it is based on false information, the little good that could come from taking the medication does not justify trying to change his mind. Thus, the doctor should accept his decision. The moral reasoner also comes to this conclusion, as can be seen in the last column of Table 5.

Table 5: Simulation results of Experiment 4.

	Autonomy	**Non-Malef**	**Benef**	**Morality**
Try Again	**-0.5**	**0**	**0.5**	**-0,26**
Accept	**0.5**	**0**	**-0.5**	**0,26**

Experiment 5

A patient with incurable cancer refuses chemotherapy that will let him live a few months longer, relatively pain free. He refuses the treatment because, ignoring the clear evidence to the contrary, he is convinced himself that he is cancer-free and does not need chemotherapy.

According to Buchanan and Brock (1989), the ethically preferable answer is to try again. The patient's less than fully autonomous decision will lead to harm (dying sooner) and denies him the chance of a longer life (a violation of the duty of beneficence), which he might later regret. The moral reasoner comes to the same conclusion, as can be seen in Table 6.

Table 6: Simulation results of Experiment 5.

	Autonomy	**Non-Malef**	**Benef**	**Morality**
Try Again	**-0.5**	**0.5**	**0.5**	**0,13**
Accept	**0.5**	**-0.5**	**-0.5**	**-0,13**

Experiment 6

Table 7: Simulation results of Experiment 6.

	Autonomy	**Non-Malef**	**Benef**	**Morality**
Try Again	**-0.5**	**0**	**1**	**0,04**
Accept	**0.5**	**0**	**-1**	**-0,04**

A patient, who has suffered repeated rejection from others due to a very large noncancerous abnormal growth on his face, refuses to have simple and safe cosmetic surgery to remove the growth. Even though this has negatively affected his career and social life, he is resigned himself to being an outcast, convinced that this is his fate in life. The doctor is convinced that his rejection of the surgery stems from depression due to his abnormality and that having the surgery could vastly improve his entire life and outlook.

The doctor should try again to convince him because so much of an improvement is at stake and his decision is less

than fully autonomous. Also here, the moral reasoner comes to the same conclusion, as can be seen in Table 7.

Discussion

The paper described a moral reasoner that combines a bottom-up structure with top-down knowledge in the form of moral duties. The reasoner estimates the influence of an action on the total amount of utility in the world by the believed contribution of the action to the following three duties: Autonomy, Non-maleficence and Beneficence. Following these three duties is represented as having three moral goals. The moral reasoner is capable of balancing between conflicting moral goals. In simulation experiments, the reasoner reached the same conclusions as expert ethicists (Buchanan & Brock, 1989).

Because the representation of goals and beliefs in the moral reasoner is very similar to the representation of beliefs and goals in the affective decision making process of Silicon Coppélia (Hoorn, Pontier & Siddiqui, 2011), the moral reasoner could easily be connected to the system. Thereby, the moral reasoning could be combined with human-like affective decision making, and the behavior of the system could be personalized for individuals.

According to Anderson, Anderson and Armen (2006), simply assigning linear weights to the moral duties is not sufficiently expressive to capture their relationships. Indeed, an extra rule had to be added to satisfy the expert data in Experiment 2. However, for all other experiments, this rule turned out not to be necessary.

Also without this rule, it would have been arguable that the moral reasoner simulates human-like moral reasoning. The analysis of the expert ethicists may not reflect the public opinion, however. Perhaps the majority of laymen would decide to question the patient's refusal to take life-saving medication. Arguably, it would not be seen as inhuman if someone did.

Even between doctors, there is no consensus about the interpretation of values and their ranking and meaning. In the work of Van Wynsberghe (2012) this differed depending on: the type of care (i.e., social vs. physical care), the task (e.g., bathing vs. lifting vs. socializing), the care-giver and their style, as well as the care-receiver and their specific needs. The same robot used in one hospital can be accepted differently depending on the ward. Workers in the post-natal ward loved the TUG-robot, while workers in the oncology ward found the robot to be rude, socially inappropriate and annoying. These workers even kicked the robot when they reached maximum frustration (Barras, 2009).

There may be doctors that feel the urge to pursue a patient to take the life-saving medication, but only choose not to do so because of ethical guidelines. It could be argued that, when health care professionals are making decisions on a strict ethical code, they are restricting their regular way of decision-making.

Further, it can be questioned whether a patient can ever be fully autonomous. According to Anderson and Anderson (2008), for a decision by a patient concerning his or her care to be considered fully autonomous, it must be intentional, based on sufficient understanding of his or her medical situation and the likely consequences of foregoing treatment. Further, the patient must be sufficiently free of external constraints (e.g., pressure by others or external circumstances, such as a lack of funds) and internal constraints (e.g., pain/discomfort, the effects of medication, irrational fears or values that are likely to change over time). Using this definition, it could be questioned whether the patient in Experiment 2 is not under the influence of external constraints (i.e., pressure from a religious leader).

Moreover, it seems that medical ethics are contradicting with the law. A fully autonomous decision of a patient wanting to commit euthanasia would be represented by the same believed contributions to following moral duties as those given in experiment 2. In the case of euthanasia, the patient also makes a fully autonomous decision that will lead to his death. However, in many countries, committing active euthanasia is illegal. In countries where euthanasia is permitted, it is usually only allowed when the patient is in hopeless suffering. By the definition of Anderson and Anderson, being in hopeless suffering would mean the patient is not free of internal constraints (i.e., pain and suffering) and therefore not capable of making fully autonomous decisions. On the other hand, in the case of hopeless suffering, it could be questioned whether one could speak of maleficence when the patient is allowed to commit euthanasia.

However, we would not like to argue against strict ethical codes in professional fields such as health care. It is important to act based on a consensus to prevent conflicts and unnecessary harm. Just as doctors restrict their 'natural' behavior by maintaining a strict ethical code, we can also let a robot restrict its behavior by acting through the same strict ethical code.

Moreover, we may well want to aim for machines that behave ethically better than human beings. Human behavior is typically far from being morally ideal, and a machine should probably have higher ethical standards (Allen et al., 2000). By matching the ethical decision-making of expert ethicists, the presented moral reasoner serves as a nice starting point in doing so.

From a cognitive science perspective, an important product of work on "machine ethics" is that new insights in ethical theory are likely to result (Anderson & Anderson, 2008). As Daniel Dennett (2006) stated, AI "makes philosophy honest". Ethics must be made computable in order to make it clear exactly how agents ought to behave in ethical dilemmas. Without a platform for testing the adequacy of a particular model of moral decision making, it can be quite easy to overlook hidden mechanisms" (Wallach, 2010).

In future research, we intend to integrate the moral reasoner with Silicon Coppélia. This could be done in various manners. Different applications might benefit from different ways of implementation.

When developing a decision-support system in the medical domain such as (Anderson, Anderson & Armen, 2006), it should have a strict ethical code. When there are conflicting moral goals, the outcome of the moral reasoning should always give the final answer on how to act. Additionally, in consult with medical ethicists and experts

from the field in which the moral reasoner will be applied, it may be necessary to add more rules to the system.

However, when developing a companion robot or virtual character that interacts with the patient, it may be more beneficial to give a bit less weight to moral reasoning. Moral goals could perhaps be treated the same as other goals that motivate the robot's behavior. In entertainment settings, we often like characters that are naughty (Konijn & Hoorn, 2005). In entertainment, morally perfect characters may even be perceived as boring. In Silicon Coppélia (Hoorn, Pontier & Siddiqui, 2011), this could be implemented by updating the affective decision making module. Morality would be added to the other influences that determine the Expected Satisfaction of an action in the decision making process. By doing so, human affective decision-making behavior could be further explored.

Acknowledgements

This study is part of the SELEMCA project within CRISP (grant number: NWO 646.000.003). We would like to thank Aimee van Wynsberghe for fruitful discussions.

References

Allen, C. Varner, G. & Zinser, J. (2000). Prolegomena to Any Future Artificial Moral Agent. *Journal of Experimental and Theoretical Artificial Intelligence*, 12, 251–61

Anderson, M., Anderson, S., & Armen, C. (2005). Toward Machine Ethics: Implementing Two Action-Based Ethical Theories. *Machine Ethics: Papers from the AAAI Fall Symposium*. Technical Report FS-05-06, Association for the Advancement of Artificial Intelligence, Menlo Park, CA

Anderson, M.; Anderson, S.; & Armen, C. (2006). MedEthEx: A Prototype Medical Ethics Advisor. *Proceedings of the Eighteenth Conference on Innovative Applications of Artificial Intelligence*. Menlo Park, CA: AAAI Press.

Anderson, M., & Anderson, S. (2007). Machine ethics: Creating an ethical intelligent agent, *AI Magazine*, 28(4), 15-26.

Anderson, M., & Anderson, S. (2008). Ethical Healthcare Agents, *Studies in Computational Intelligence*, 107, Springer.

Anderson, M., & Anderson, S. (2010), Robot be Good, *Scientific American*, October 2010, 72-77.

Banks, M.R., Willoughby, L.M., and Banks, W.A. (2008). Animal-Assisted Therapy and Loneliness in Nursing Homes: Use of Robotic versus Living Dogs. *Journal of the American Medical Directors Association*, 9, 173-177

Barras C. (2009) Useful, loveable and unbelievably annoying. *The New Scientist*, 22-23.

Buchanan, A.E. and Brock, D.W. 1989. *Deciding for Others: The Ethics of Surrogate Decision Making*, Cambridge University Press.

Dennett, D. (2006). *Computers as Prostheses for the Imagination*. Invited talk presented at the International Computers and Philosophy Conference, Laval, France, May 3.

Gillon R. (1994) Medical ethics: four principles plus attention to scope. *BMJ*. 309(6948), 184–188.

Guarini, M. (2006). Particularism and the Classification and Reclassification of Moral Cases. *IEEE Intelligent Systems*, 21(4), 22–28.

Hoorn, J.F., Pontier, M.A., & Siddiqui, G.F., (2011). Coppélius' Concoction: Similarity and Complementarity Among Three Affect-related Agent Models. Cognitive Systems Research Journal, in press.

Kidd, C., Taggart, W., and Turkle, S. (2006). A Social Robot to Encourage Social Interaction among the Elderly. *Proceedings of IEEE ICRA*, 3972-3976

Konijn, E.A., & Hoorn, J.F. (2005). Some like it bad. Testing a model for perceiving and experiencing fictional characters. *Media Psychology*, 7(2), 107-144.

Marti, P. Bacigalupo, M., Giusti, L., and Mennecozzi, C. (2006). Socially Assistive Robotics in the Treatment of Behavioral and Psychological Symptoms of Dementia. *Proceedings of BioRob*, 438-488.

Picard R (1997) Affective computing. MIT Press, Cambridge, MA

Robins, B., Dautenhahn, K., Boekhorst, R.T., and Billard, A. (2005). Robotic Assistants in Therapy and Education of Children with Autism: Can a Small Humanoid Robot Help Encourage Social Interaction Skills? *Journal of Universal Access in the Information Society*. 4, 105-120.

Ross, W. D. (1930). The Right and the Good. Oxford: Clarendon Press.

Rzepka, R., & Araki, K. (2005). What Could Statistics Do for Ethics? The Idea of a Common Sense Processing-Based Safety Valve. In *Machine Ethics: Papers from the AAAI Fall Symposium*. Technical Report FS-05-06, Association for the Advancement of Artificial Intelligence, Menlo Park, CA.

Super, D.E. (1973). The work values inventory. In D.G. Zytowski (Ed.), *Contemporary approaches to interest measurement.* Minneapolis: University of Minnesota Press.

Tronto, J. (1993). *Moral Boundaries: a political argument for an ethic of care*. Routledge, New York.

Van Wynsberghe, A. (2012). Designing Robots for Care; Care Centered Value-Sensitive Design. *Journal of Science and Engineering Ethics*, in press

Wada, K., and Shibata, T. (2009). Social Effects of Robot Therapy in a Care House , *JACIII*, 13, 386-392

Wallach, W. (2010). Robot minds and human ethics: The need for a comprehensive model of moral decision making. *Ethics and Information Technology*, 12(3), 243-250.

Wallach, W., Franklin, S. & Allen, C. (2010). A Conceptual and Computational Model of Moral Decision Making in human and Artificial Agents. *Topics in Cognitive Science*, 2, 454–485.

Wallach, W., Allen, C., & Smit, I. (2008). Machine morality: Bottom-up and top-down approaches for modelling human moral faculties. *AI and Society*, 22(4), 565–582.

WHO (2010) *Health topics: Ageing*. Available from: http://www.who.int/topics/ageing/en/